CLAIMS:

1. A system for streaming a software application to a plurality of clients comprising:

a principal server having the software stored thereon as a plurality of blocks and comprising a principal predictive streaming application configured to predict blocks which will be required by devices connected to the principal server, and a principal streaming communication manager configured to transmit predicted blocks to designated devices connected to the principal server and service requests for blocks issued from downstream devices;

at least one intermediate server connected between the principal server and the plurality of clients, each intermediate server connected to at least one upstream device and at least one downstream device and comprising a cache, a respective intermediate predictive streaming application configured to predict blocks which will be required by connected downstream devices, and a respective intermediate streaming communication manager;

each respective intermediate streaming communication manager configured to (a) transmit predicted blocks to designated downstream devices, (b) service requests for blocks issued from downstream devices, (c) cache blocks received from connected upstream devices, and (d) issue requests for a particular block to an upstream device when the particular block is needed for transmission to a downstream device and is not present in the cache;

wherein each device comprises one of an intermediate server and a client.

2. The system of claim 1, wherein the intermediate streaming communication manager is further configured to, in response to an indication that a cache purge is required,

select at least one block to purge in accordance with a determination of a cost to replace particular blocks in the cache.

3. The system of claim 2, wherein the intermediate communication streaming manager is further configured to determine the cost to replace particular blocks in the cache with reference to cached contents at connected devices.

4. The system of claim 3 wherein the intermediate communication streaming manager is further configured to broadcast to at least some of the connected devices indications of caching and purging events.

5. The system of claim 4, wherein the intermediate communication streaming manager is configured to broadcast caching and purging event indications to direct descendant and direct ancestor devices.

6. The system of claim 1, wherein the intermediate streaming communication manager is further configured to:

generate a reference value for each block in the associated cache related to a cost to replace the particular block in the cache; and

upon a determination that a cache purge is required, select at least one block to purge from a set of blocks having a reference value exceeding a predefined threshold.

7. The system of claim 6, wherein the cost is determined with reference to cached contents at connected devices.

8. The system of claim 7, wherein the intermediate streaming communication manager is further configured to recalculate the reference values for blocks in the associated cache upon a receipt of a broadcast from a connected device indicating a change in cache contents at that connected device.

9. The system of claim 8, wherein the intermediate streaming communication manager is further configured to broadcast to at least some of the connected devices indications of caching and purging events.

10. The system of claim 6, wherein the cost for a respective block is determined with reference to at least one of :

a block size;

a cost in CPU tasks to stream the respective block to the intermediate server from a connected device which is an alternative source of the respective block;

quality of transmission line to the alternative source of the respective block;

type of transmission line to the alternative source of the respective block;

cost to store and maintain the block at the particular intermediate server;

distance in network nodes to the alternative source of the respective block; and

frequency of use of the respective block.

11. The system of claim 1, wherein the intermediate predictive streaming application is configured to predict blocks which will be required by immediate downstream descendant devices.

12. The system of claim 1, wherein the intermediate streaming communication manager is configured to request blocks from upstream devices in accordance with the prediction of blocks which will be required by downstream devices.

13. A server for use in a system for streaming a software application to a plurality of clients comprising:

a cache;

a predictive streaming application configured to predict blocks which will be required by connected downstream devices; and

a streaming communication manager configured to (a) transmit predicted blocks to designated downstream devices, (b) service requests for blocks issued from downstream devices, (c) cache blocks received from connected upstream devices, and (d) issue requests for a particular block to an upstream device when the particular block is needed for transmission to a downstream device and is not present in the cache;

wherein each device comprises one of a server and a client.

14. The system of claim 13, wherein the streaming communication manager is configured to request blocks from an upstream device in accordance with the prediction of blocks which will be required by a connected downstream device.

15. The system of claim 13, wherein the streaming communication manager is further configured to, in response to an indication that a cache purge is required, select at least

one block to purge in accordance with a determination of a cost to replace particular blocks in the cache.

16. The system of claim 14, wherein the communication streaming manager determines the cost to replace particular blocks in the cache with reference to cached contents at devices connected to the server.

17. The system of claim 15 wherein the communication streaming manager is configured to broadcast to at least some devices connected to the server indications of caching and purging events.

18. The system of claim 16, wherein devices connected to the server are organized in a tree configuration and the communication streaming manager is configured to broadcast caching and purging event indications to direct descendant and direct ancestor devices connected to the server.

19. The system of claim 13, wherein the streaming communication manager is further configured to:

generate a reference value for each block in the associated cache related to a cost to replace the particular block in the cache; and

upon a determination that a cache purge is required, select at least one block to purge from a set of blocks having a reference value exceeding a predefined threshold.

20. The system of claim 18, wherein the cost is determined with reference to cached contents at devices connected to the server.

21. The system of claim 19, wherein the streaming communication manager is further configured to recalculate the reference values for blocks in the associated cache upon a receipt of a broadcast from a device connected to the server indicating a change in cache contents at that connected device.

22. The system of claim 20, wherein the streaming communication manager is further configured to broadcast to at least some devices connected to the server indications of caching and purging events.

23. The system of claim 18, wherein the cost for a respective block is determined with reference to at least one of :

a block size;

a cost in CPU tasks to stream the respective block to the server from a connected device which is an alternative source of the respective block;

transmission line quality to the alternative source of the respective block;

transmission line type to the alternative source of the respective block;

cost to store and maintain the block at the particular intermediate server;

distance in network nodes to the alternative source of the respective block from the intermediate server; and

frequency of use of the respective block.

24. The system of claim 13 wherein the predictive streaming application is configured to predict blocks which will be required by immediate downstream descendant devices connected to the server.

25. In a system for streaming a software application as blocks from a principal server to at least one client having at least one intermediate server between the principal server and the client, each intermediate server connected to at least one upstream device and at least one downstream device, each device comprising one of the principal server, a client, and another intermediate server, a method for improving the deliver of the software application comprising the steps of:

predicting at the intermediate server blocks which will be required by a downstream device;

transmitting predicted blocks from the intermediate server to a designated downstream device;

caching blocks at the intermediate server received from an upstream device in a cache;

receiving requests at the intermediate server from a particular downstream device for a particular block; and

issuing requests for the particular block from the intermediate server to the upstream device when the requested particular block is not present in the intermediate server cache; and

transmitting the particular block from intermediate server to the particular downstream device.

26. The method of claim 25, further comprising the step of issuing requests from the intermediate server to the upstream device for blocks which have been predicted to be required by a connected downstream device and are not in the intermediate server cache.

27. The method of claim 25, further comprising the step of:

determining the cost to replace particular blocks in the intermediate server; and

in response to an indication that a cache purge is required at the intermediate server, selecting at least one block to purge from the intermediate server cache in accordance with the determined cost.

28. The method of claim 26, wherein the step of determining the cost comprises considering cache contents at devices connected to the intermediate server.

29. The method of claim 27, further comprising the step of broadcasting from the intermediate server indications of caching and purging events.

30. The method of claim 25, further comprising the steps of:

generating a reference value for each block in the intermediate server cache, the reference value related to a cost to replace the particular block in the cache; and

upon a determination that a cache purge is required at the intermediate server, selecting at least one block to purge from a set of blocks having a reference value exceeding a predefined threshold.

31. The method of claim 30, wherein the cost is determined with reference to cached contents at devices connected to the intermediate server.

32. The method of claim 31, further comprising the step of recalculating the reference values for blocks in the intermediate server cache upon a receipt at the intermediate server of broadcast from a connected device indicating a change in cache contents at that connected device.

33. The method claim 32, further comprising the step of broadcasting from the intermediate server to at least some devices connected to the server indications of caching and purging events at the intermediate server.

34. The method of claim 30, wherein the generated cost is determined with reference to at least one of :

a block size;

a cost in CPU tasks to stream the respective block to the intermediate server from a connected device which is an alternative source of the respective block;

transmission line quality to the alternative source of the respective block;

transmission line type to the alternative source of the respective block;

cost to store and maintain the block at the intermediate server;

distance in network nodes to the alternative source of the respective block from the intermediate server; and

frequency of use of the respective block.

35. A computer program product for use a system for streaming a software application as blocks from a principal server to at least one client having at least one intermediate server between the principal server and the client, each intermediate server connected to at least one upstream device and at least one downstream device, each device comprising one of the principal server, a client, and another intermediate server, the computer program product comprising computer code to configure an intermediate server to:

predict blocks which will be required by a downstream device;

transmit predicted blocks to a designated downstream device;

cache blocks received from an upstream device in a cache;

receive requests from a particular downstream device for a particular block; and

issue requests for the particular block to the upstream device when the requested particular block is not present in the cache; and

transmit the particular block to the particular downstream device.

36. The computer program product of claim 35, further comprising computer code to configure the intermediate server to issue requests to the upstream device for blocks which have been predicted to be required by a connected downstream device and are not in the cache.

37. The computer program product of claim 35, further comprising computer code to configure the intermediate server to:

determine the cost to replace particular blocks in the cache; and

when a cache purge required, select at least one block to purge from the cache in accordance with the determined cost.

38. The computer program product of claim 36, further comprising computer code to determine the cost with reference to cache contents devices connected to the intermediate server.

39. The computer program product of claim 37, further comprising computer code to configure the intermediate server to broadcast indications of caching and purging events.

40. The computer program product of claim 35, further comprising computer code to configure the intermediate server to:

generate a reference value for each block in the cache, the reference value related to a cost to replace the particular block in the cache; and

upon a determination that a cache purge is required at the intermediate server, select at least one block to purge from a set of blocks having a reference value exceeding a predefined threshold.

41. The computer program product of claim 40, wherein the cost is determined with reference to cached contents at devices connected to the intermediate server.

42. The computer program product of claim 35, further comprising computer code to configure the intermediate server to recalculate the reference values for blocks in the

cache upon a receipt at the intermediate server of broadcast from a connected device indicating a change in respective cache contents at that connected device.

43. The computer program product of claim 42, further comprising computer code to configure the intermediate server to broadcast to at least some devices connected to the server indications of caching and purging events.

44. The computer program product of claim 40, wherein the cost is determined with reference to at least one of :

a block size;

a cost in CPU tasks to stream the respective block to the intermediate server from a connected device which is an alternative source of the respective block;

transmission line quality to the alternative source of the respective block;

transmission line type to the alternative source of the respective block;

cost to store and maintain the block at the intermediate server;

distance in network nodes to the alternative source of the respective block from the intermediate server; and

frequency of use of the respective block.